基于最大信息传递熵的 ICS 因果关系建模 *

张仁斌 [a, b, c†]，曹宗泽 [a]，吴克伟 [a]

(合肥工业大学 a. 计算机与信息学院; b. 大数据知识工程教育部重点实验室; c. 工业安全与应急技术安徽省重点实验室，合肥 230601)

摘　要：针对传统因果关系算法难以准确分析含大量噪声的非线性数据的问题进行了研究，提出基于最大信息传递熵的因果关系建模算法。首先，利用最大信息系数对非线性数据的时序趋势间的关联度进行检测，弱化噪声对变量间相关性的影响；然后根据筛选因子剔除弱相关变量，并通过随机经验估值计算强关联变量间的传递熵，以减少传递熵的计算量；最后，传递熵确定因果关系方向，形成支持链路溯源的单向因果网络。利用经典化工过程数据集对该算法进行测试分析，实验结果表明，相比于现有因果关系建模算法，该算法可定位异常变量，对 12 维以上的高维数据建模的稳定性高于 85%，因果关系的准确率可达 83.33%，实际建模效果优于对比算法，可用于工业控制系统异常检测定位。

关键词：工业控制系统；因果关系建模；最大信息传递熵；链路溯源；异常定位

中图分类号：TP391.4　　doi: 10.19734/j.issn.1001-3695.2020.01.0033

Ics causality modeling based on maximum information transfer entropy

Zhang Renbin[a, b, c†], Cao Zongze[a], Wu Kewei[a]

(*a. School of Computer Science & Information Engineering*, *b. Key Laboratory of Knowledge Engineering with Big Data*, *c. Anhui Province Key Laboratory of Industry Safety & Emergency Technology*, *Hefei University of Technology*, *Hefei Anhui* 230601, *China*)

Abstract: This paper developed a causality modeling algorithm based on maximum information transfer entropy to solve the problem that traditional causality algorithms were difficult to accurately analyze non-linear data with a lot of noise. First, used the maximum information coefficient to detect the correlation between time series trends of non-linear data. Weaken the effect of noise on the correlation between variables. Secondly, eliminated weakly related variables based on screening factors. Calculated the transfer entropy between strong correlations using stochastic empirical valuation. Thereby reducing the calculation amount of transfer entropy. Finally, transfer entropy determined causal direction. Formed a one-way causal network that supports link traceability. Test analysis of the algorithm using classic chemical process data sets. Test results show that, compared to existing algorithms, this algorithm can locate abnormal variables. The stability of this algorithm for modeling high-dimensional data of more than 12 dimensions is higher than 85%, and the accuracy rate of causality can reach 83.33%. The actual modeling effect of this algorithm is better than the comparison algorithms, and it can detect and locate industrial control system abnormalities.

Key words: industrial control system; causality modeling; maximum information transfer entropy; link traceability; anomaly location

0　引言

工业控制系统(industrial control system，ICS)在工业生产控制中的广泛应用，加快了生产自动化与智能化的发展进程，也引发了很多安全问题[1]。ICS 中各器件普遍存在交错复杂的依赖关系，使得其物理过程中设备节点之间的相互影响变得难以分析，导致难以准确定位异常。由因果关系形成的因果图[2]可以代表信息的传播方向，允许分析人员遵循关系链路追踪异常源头[3]，因此可以使用因果图反映 ICS 物理层设备间的依赖关系并为高效的安防提供理论指导。

目前，一种典型的因果关系研究方法是格兰杰因果关系(granger causality)。Ma L 等人[4]提出了一种基于神经网络架构的格兰杰因果分析方法，用于 KPI (key performance indicator) 定向故障的传播路径识别。Kathari S 等人[5]针对多变量平稳线性动态过程提出了一种有效重构加权格兰杰因果网络的系统方法。然而，传统的格兰杰因果关系是基于系统过程的自回归模型[6]，适用于线性多变量过程，对于非线性因果关系不敏感。

贝叶斯网络是研究非线性数据间因果关系的经典方法之一。Zhang Q 等人[7]提出了一种用于动态风险评估的模糊概率贝叶斯网络，并嵌入了噪声证据过滤器，以减少噪声数据对算法的影响，但过滤噪声也造成了一定的信息损失，并且传统贝叶斯网络更适用于离散数据。为处理连续数据，YANG Jing 等人[8]提出了基于 PCB (partial correlation-based)算法的连续贝叶斯网络模型，但该模型对非线性结构数据的因果分析效果不佳。为将离散数据建模扩展为连续数据建模并且能够挖掘非线性数据中的潜在关系，曾千千等人[9]利用最大信息系数(maximum information coefficient, MIC)搭建基础关系

收稿日期：2020-01-11；**修回日期：**2020-03-13　**基金项目：**国家重点研发计划专项资助项目(2016YFC0801804，2016YFC0801405)；中央高校基本科研业务费专项资金资助项目(PA2019GDPK0074)

作者简介：张仁斌(1971-)，男(通信作者)，湖北汉川人，副教授，博士，主要研究方向为工业互联网安全(zhangrb@hfut.edu.cn)；曹宗泽(1994-)，男，天津人，硕士研究生，主要研究方向为工业互联网安全(446542574@qq.com)；吴克伟(1984-)，男，安徽合肥人，副教授，博士，主要研究方向为计算机视觉，人工智能，模式识别.

网络框架，通过贪婪算法对 MIC 构造的贝叶斯网络结构进行局部优化，通过整合局部最优解生成最终的网络结构。但该方法受限于贪婪算法的局部最优特性，无法保证每次结果均为全局最优，导致建模结果不稳定。

相较于格兰杰因果性，传递熵(transfer entropy，TE)更加精确，从中得出的因果关系图在视觉上更易于解释 [10]。 Shi D 等人[11]针对传感器测量序列引入了基于传递熵 TE 的因果对策，以数据驱动的方式对其进行评估而无须依赖基础动态系统的模型。但传递熵 TE 的计算复杂度高，存在效率不高的问题。Su J 等人[12]通过比较连续变量的相应阈值生成离散报警序列，降低了传递熵 TE 的计算代价，提出了一种基于传递熵 TE 和修正互信息的混合方法检测变量之间直接和间接因果关系，但其算法所得结果是双向因果关系，不能完成有限长度的链路溯源及异常定位；同时将连续数据离散化为报警序列也会损失一定的原始信息。

综上所述，目前，在工业控制系统中进行因果关系建模时，主要存在以下几个问题：

a)实际生产数据往往为连续非线性且伴随大量噪声，离散化数据或滤波处理等预处理常以损失信息为代价从而影响后续精准分析。既能分析连续非线性数据又无须过滤噪声成为因果关系分析的关键问题。

b)含有回路或结果不稳定的因果关系网络，不适用于异常点排查，对同一对象多次建模的结果差异较大将给后续分析带来干扰、影响结果准确性。因此需要建模后的因果网络能够形成单向稳定链路的算法。

针对上述问题，本文提出了基于最大信息传递熵的因果关系建模算法，即 MITE-CM (maximum information transfer entropy causal modeling)。本文算法利用 MIC 建立相关性网络框架，以获取在时序趋势上具有强相关性的连续非线性数据，减少噪声对计算信息熵的影响；通过传递熵 TE 反映强相关变量间的信息传递方向，形成有利于链路溯源的单向无回路网络，并解决因果网络结构不稳定的问题。算法通过筛选机制过滤弱相关关系，减少传递熵 TE 部分的计算量。

1　基于最大信息传递熵的因果关系建模算法

1.1　MITE-CM 算法

工业控制系统中的复杂关系非单一的函数关系且数据中包含大量噪声，需要普适性强、鲁棒性高的算法。受文献[13]启发，本文综合两种相关性分析方法，同时解决问题 a)和问题 b)。MITE-CM 利用 MIC 指标衡量两个变量 X 和 Y 之间线性或非线性的关联程度。MIC 的普适性决定算法在样本量足够大时能够捕获多种关联，而非限定于单一的函数类型。MIC 的公平性保证其在样本量足够大时，能为噪声程度相似的不同种相关关系给出相近的系数。例如，对于充满相同噪声的线性关系和正弦关系，MIC 能给出相近的相关系数。因此无须对全体数据集特别进行滤波等除噪预处理，可直接计算原始数据。相比于 FPBN[7](fuzzy probability Bayesian network)等贝叶斯网络方法，MITE-CM 算法允许直接处理连续数据，同时对正常噪声有良好的鲁棒性，可直接分析 ICS 数据间复杂多样的关联情况，可以解决问题 a)。

但由于 MIC 的对称性，仅由 MIC 形成的相关性网络不具备方向性。为形成有向图，MIC-GA 算法[9]利用贪婪搜索将 MIC 网络扩展为有向因果图，却未证明其局部最优解即为系统实际因果关系，并且无法保证局部最优即为全局最优，使得准确性相对较弱。传递熵根据其不对称性建立驱动和响应间的因果关系，不需要系统模型机理的先验知识。MITE-CM 算法将传递熵 TE 与 MIC 网络融合得到相对稳定准确的因果关系。首先对序列 X、Y 进行网格化以此得出 X、Y 之间的最大互信息，并将其归一化进而分析两序列整体趋势间的相关性；随后计算 MIC 网络中的强相关关系($mic_{i,j} \geq th$)间的传递熵，将具有较大传递熵的方向视为信息流动方向，使不具备方向性的相关性关系转换为单向的因果关系，如图 1 所示。

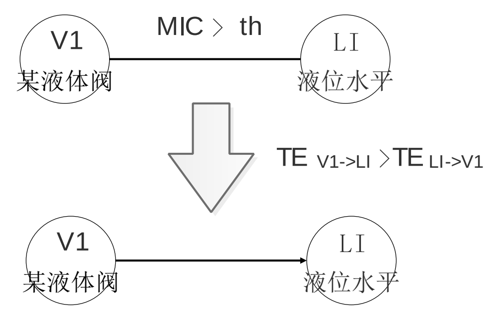


图 1　相关性关系转换为因果关系

Fig. 1　Correlation turns into causation

对于前文归纳问题 b)，MITE-CM 算法以变量间的时序趋势相关性为关键特征，保留了具有强时序相关性的 MIC 关系，使得因果图的基本框架不易变动。此外，MITE-CM 以传递熵的规则判定因果关系的方向，仅在双方信息熵大小极为接近时其所得方向才有可能改变，确保系统正常情况下建模结果差异性小。换而言之，若 MITE-CM 的两次建模结果出现框架差异，则系统出现异常，细节将在实验与分析中详述。而 MIC-GA 算法中的贪婪策略常因局部最优解不同而变更网络框架，使得多次实验结果的差异程度较大，为后续的进一步分析带来诸多干扰。与 MIC-GA 算法的随机性对比将在稳定性实验中做进一步说明。MITE-CM 算法伪码如下所示。

算法 1　MITE–CM 算法

输入： 数据集 ***X***

输出： 因果关系矩阵 ***E***

```
a) let C be a new N*N matrix.
b) for i = 1 to N
c)      for j = 1 to N
d)          if j == i
e)              c_{i,j} ← 0 ;
f)          else
g)              c_{i,j} ← MINE(X_{.i}, X_{.j}).mic ;      /* MIC 系数*/
h) let E be a new N*N matrix.
i) for i = 1 to N
j)      for j = i to N
k)          if |c_{i,j}| > th  /*筛选阈值*/
l)              (t_{i→j}, t_{j→i}) ← TE(R_EVM(X_{·i}, X_{·j}, len));
/* 计算传递熵 TE ，计算传递熵所需的概率密度由 R_EVM 算法得出*/
m)              if t_{i→j} = 0 and t_{j→i} = 0
n)                  c_{i,j} , c_{j,i} ← 0 ;
o)              else if t_{i→j} − t_{j→i}  ⩾ 0
p)                  c_{i,j} ← 1 ;
q)              else
r)                  c_{j,i} ← 1 ;
s) return E;
```

1.2　R_EVM 算法

工业控制系统中数据的真实概率分布通常未知，在计算传递熵 TE 前须近似估计变量的概率分布。相比与传统概率密度估计方法，属于非参数概率密度估计的经验估值法能处理任意形式的概率分布且不需要作出假设。本文利用经验估值法计算简单、与总体分布相关的特点分析数据整体趋势间因果关系。为减少计算量，本文将经验估值法改进为随机经验估值法 R_EVM(random empirical valuation method)，首先获取待测序列 X、Y 的长度 L 及值域 R；其次将值域 R 分割

为 p 个子区间；随后分别抽取待测序列 X、Y 的 len 个数据形成新的序列 X'、Y'；最后分别统计 X'、Y'中数据落入各个自区间中的个数，以其占比代表该区间的概率密度函数值，并存储于概率密度函数数组 P 中。为保证估值结果准确，确保随机选取的数据遍及待测序列的绝大部分，R_EVM 算法设定 len 取值不小于输入序列长度 L 的一半，即 $\text{len} \geq L/2$ 。基于 R_EVM 的传递熵 TE 统计区间分布探查变量宏观趋势间的因果性，对系统噪声不敏感，同样满足前文归纳问题 a)中算法无须过滤噪声的要求。R_EVM 算法计算传递熵公式中部分概率密度的伪码如下：

算法 2　R_EVM 算法

```
输入：序列 x 和 y,区间数量 p,采样总数 len.
输出：p(x_t,y_t).
a) △1 = (X_max - X_min) / (2 * p); △2 = (Y_max - Y_min) / (2 * p) ;
b) pointer[]←random(len) ;//随机生成 len 个下标
c) X_t[] ← X(pointer) ; Y_t[] ← Y(pointer) ; Y_t+1[] ← Y(pointer +1) ; X_t+1[]←X(pointer +1);
d) Lx[]←X_min+△1 : p : X_max-△1; Ly[]←Y_min+△2 : p : Y_max-△2;   /*分割区间,每区间间隔为△ */
e) stat=zeros(p, p, 3);   /*三维全零矩阵,统计数据在各区间的分布情况*/
f) for i=1:p
g)     for j=1:p
h)         count←0 ;
i)         for k=1:len
j)             if (Lx(i)- △1) ≤ X_t(k)≤(Lx(i)+ △1) and (Ly(j)- △2) ≤ Y_t(k)≤(Ly(j)+ △2)
k)                 count ++ ;
l)         stat(i,j,3)←count ;
m) p(x_t,y_t) = stat(:,:,3)/sum(sum(stat(:,:,3))) ;
n) return p(x_t,y_t) ;
```

由伪码可知，在计算传递熵公式的联合概率 $p(x_t, y_t)$ 时，其复杂度已经达到 $O(n^3)$。序列 X'、Y'较短所分区间过多，导致计算效率不高；而序列 X'、Y'较长所分区间较少，易使结果精度下降。因此，p^2 与 len 应处于相同量级，即 $p/\text{len} \approx 0.05$。通常情况下，抽样长度 $len \approx 1000$ 时，区间数 $p \approx 50$。此时，相比于传统经验估值法，R_EVM 算法理论上可减少 $\frac{L \times p^2}{2} \approx 1.25 \times 10^6$ 次计算。相比于 BAS-TE 算法[11]将数据离散化为五类定值并计算传递熵，R_EVM 中保留了原始数据的多样性，远多于五种分类的区间数 p 可捕获更多的信息，使计算结果理论上更为精确，算法准确性对比将在精确性实验中讨论。此外，R_EVM 存在一定随机性，但在面向整体分布分析数据的目的下该随机性在可接受范围内，具体情况将在稳定性实验中的 len/L 取值分析中说明。

1.3　筛选因子

计算变量概率密度分布的复杂度较高，随着网络规模的不断增大，R_EVM 的调用次数也呈现平方式的增长。因此，算法在计算传递熵 TE 前通过由筛选因子 α 计算出的阈值 th 过滤弱相关性的变量，再次减少 R_EVM 的计算量。本文将筛选因子 α 设为第二位有效数字的计数单位，筛选阈值 th 的计算公式如式(1)所示。

$$\text{th} = \begin{cases} \left[\dfrac{ave}{\alpha}\right] \times \alpha, & ave \geq 1 \\ \dfrac{[ave \times \alpha]}{\alpha}, & ave < 1 \end{cases} \tag{1}$$

其中 α 为筛选因子；th 为筛选阈值；$[\cdot]$ 为四舍五入取整运算；ave 为非零 $mic_{i,j}$ 的均值，其计算公式如式(2)所示。

$$ave = \frac{\sum mic_{i,j}}{n}, mic_{i,j} \neq 0 \tag{2}$$

其中，n 为非零 $mic_{i,j}$ 总数。

若随机变量 X 和 Y 之间的 MIC 值大于筛选阈值 th，则认为这两个变量之间存在一条边，MITE-CM 算法随后计算这两个变量之间的传递熵 TE。筛选因子使 MIC 框架中保留了主要因果关系，同时减少弱相关性对整体网络带来的干扰，进而减少后续传递熵 TE 部分的计算。

2　实验与分析

2.1　经典化工过程

田纳西-伊斯曼过程[14] (Tennessee-Eastman process, TEP)是经典的化工过程模型，如图 2 所示，它模拟了连续过程所面临的大多数挑战[15]，被广泛应用于工厂控制策略设计、多变量控制、稳态与动态优化、预测控制、自适应控制、非线性控制等领域的研究[16,17]。本文所使用的 TEP 数据为 Kaspersky 仿真的 TEP 正常数据及其攻击测试后的异常数据[18]。

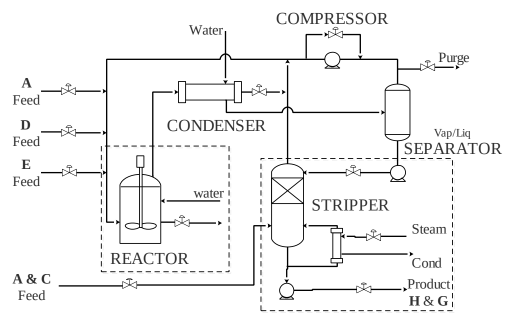


图 2　田纳西—伊斯曼过程图

Fig. 2　Tennessee—Eastman process diagram

为充分说明本文算法对 ICS 物理过程数据的因果分析通用性及异常定位的普适性，本章分别测试 TEP 中的 DDoS 攻击异常数据和完整性攻击异常数据。两种异常分别来自 TEP 过程中的汽提塔模型和反应罐模型，如图 1 虚线框所示，相关变量如表 1 所示，实验中的筛选阈值 th 均为 0.1。

表 1　攻击异常相关变量

Tab. 1　Attacks the exception correlation variable

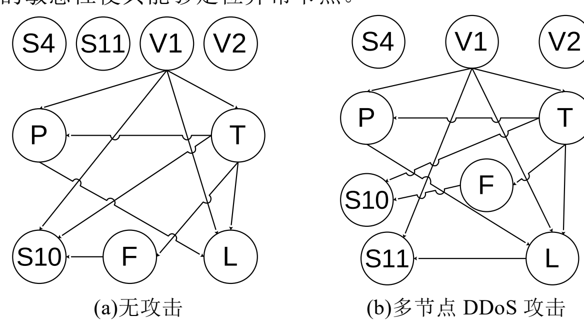
符号	名称	符号	名称
S4	A+C Feed	V1	Stripper Liquid Product Flow
S10	Sep Underflow	V2	Stripper Steam Flow
S11	Stripper Underflow	RT	Recator Temperature
P	Stripper Pressure	RF	Recator Feed
T	Stripper Temperature	RP	Recator Proseeure
F	Steam Flow	RL	Recator Level
L	Stripper Level	ReF	Recycle Flow

2.2　汽提塔因果分析及 DDoS 攻击检测

通过对汽提塔变量进行计算得到的因果关系网络如图 3(a)所示。汽提塔内部的因果关系从控制阀 *V1* 展开，并且与汽提塔的压力指示器 *P*、温度指示器 *T*、进料流 *S10* 和液位水平 *L* 产生了直接因果关系。同时，压力指示器 *P*、温度指示器 *T* 和汽提塔下溢流量计 *F* 又分别延伸出各自的因果关系，从而形成了许多类似“产量阀 *V1*→汽提塔内压力 *P*→汽提塔内液位水平 *L*”的级联因果关系。在系统正常运行时，控制阀 *V2*、流量 *S4* 和 *S11* 与其他节点的数据波形不具备明显相似性($mic_{i,j} < th$)，因此不存在因果关系。其余因果关系均符合模型正常运行时的生产逻辑，算法因果建模具备合理性。

本文算法测试 DDoS 攻击后一场数据的因果网络如图 3(b)所示。对比攻击前后，*S11* 从与其他节点没有因果关系变为与 *V1* 和 *L* 存在直接因果关系，新增了因果关系“*V1*→*S11*”

和“$L \rightarrow S11$”，而因果关系“$V1 \rightarrow S10$”也随之消失，其余的因果关系保持不变。由此推断，主要异常节点为 *V1*、*S11* 和 *L*，节点 *S10* 为受影响节点，由异常节点溯源可得异常链路 Track，分析结果如表 2 所示。Kaspersky 数据集显示，节点 *S11*、*L* 和 *V1* 中包含异常数据，与推测结果一致。说明因果关系的变动可以作为评判异常的标准之一，而本算法对因果关系变动的敏感性使其能够定位异常节点。



(a)无攻击 (b)多节点 DDoS 攻击

图 3 汽提塔因果关系

Fig. 3 Stripper causality

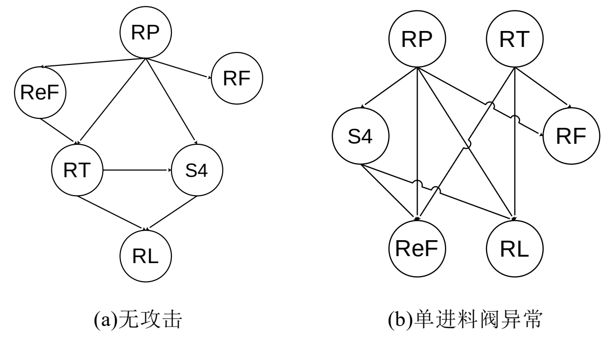
表 2 汽提塔异常情况

Tab. 2 Stripper abnormal condition

集合	节点元素
Abnormal_set	*S11;L;V1*
Impact_set	*S10*
Track	$V1 \rightarrow S11; V1 \rightarrow P \rightarrow L \rightarrow S11;$ $V1 \rightarrow T \rightarrow L \rightarrow S11; V1 \rightarrow L \rightarrow S11;$ $V1 \rightarrow P \rightarrow L; V1 \rightarrow T \rightarrow L; V1 \rightarrow L; V1$

2.3 反应罐因果分析及异常溯源对照

再次对 TEP 中反应罐变量测试，系统正常运行下反应罐变量存在的因果关系，结果如图 4(a)所示。



(a)无攻击 (b)单进料阀异常

图 4 反应罐因果关系

Fig. 4 Reaction tank causality

算法得出以 D 进料阀为初始异常的因果关系网络，如图 4(b)所示。对比攻击前后因果关系，新增因果关系“$S4 \rightarrow REF$”“$RT \rightarrow REF$”“$RT \rightarrow RF$”，缺失因果关系“$RP \rightarrow RT$”“$REF \rightarrow RT$”“$RT \rightarrow S4$”，以新增关系的节点为异常点，以消失关系的上游节点为受影响点，二者交集同样视为异常点。本次以因果关系的增加与消失(框架差异)作为异常的传播路径 new_Track，统计结果如表 3 所示。生成的新型异常路径与 MFM-SDG 算法[19]对反应罐异常分析得到的路径“$RF \rightarrow RT \rightarrow RP$”中所含节点一致，但方向相反。这是由于本算法考虑信息的“流动”方向，通常下游节点具有更高的信息熵，而信息熵较高的节点多为工艺流程顺序中的上游节点。因此将路径倒置后，两算法溯源结果相同，再次证实 MITE-CM 对因果关系异常敏感，算法准确度较高。由于 MFM-SDG 算法将多个节点(*MV1~MV6*)设为初始异常，而 Kaspersky 此次攻击测试的数据中仅将 D 进料阀节点(*MV1*)设为异常，因此本算法所得故障路径的逆序集合包含于 MFM-SDG 算法异常路径集合。

表 3 反应罐异常情况

Tab. 3 Reaction tank anomaly

集合	节点元素
Abnormal_set	*RT;S4;RF*
Impact_set	*REF;RP*
new_Track	$RT \rightarrow S4; REF \rightarrow RT;$ $RP \rightarrow RT \rightarrow RF$

综上所述，在 TEP 中使用不同攻击后的异常数据测试算法，实验结果表明，根据 MITE-CM 算法在不同时刻建立的因果关系模型判断系统中的因果关系变更，可定位工业控制系统多种异常节点。

3 算法对比分析

将本文算法实验结果分别与 MIC-GA(Maximum information coefficient - Greedy Algorithm) 算法[9]、BAS-TE(Binary Alarm Sequence - Transfer Entropy)算法[11]和 TE-CMI(Transfer Entropy - Conditional Mutual Information)算法[20]的实验结果进行对比，相关变量如表 4 所示，MEAS 和 MV 分别表示变量类型为测量变量和控制变量。

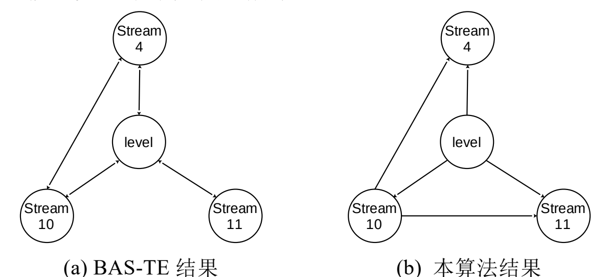
表 4 部分 TEP 过程变量

Tab. 4 Partial TEP process variables

符号	名称	单位	类别
Stream 4	A+C Feed	kscmh	MEAS
Stream 6	Reactor Feed	kscmh	MEAS
Stream 8	Recycle Flow	kscmh	MEAS
Stream 10	Sep Underflow	m³/h	MEAS
Stream 11	Stripper Underflow	m³/h	MEAS
Valve 1	A Feed Flow	%	MV
Valve 9	Purge Flow	%	MV
level	Stripper Level	%	MEAS

3.1 功能性分析

本文算法与 BAS-TE 算法[11]的实验结果进行了对比，当本文算法的筛选阈值 *th* 为 0.060 时，结果与 BAS-TE 算法结果近似。实验结果如图 5 所示。



(a) BAS-TE 结果 (b) 本算法结果

图 5 因果关系对比图

Fig. 5 Causal correlation diagram

由图 5(a)可知，BAS-TE 算法生成的因果关系网络产生了环状链路。从定位异常节点的角度来看，本文算法生成的无回路因果网络更有利于揭示 ICS 物理过程数据之间的因果关系，并根据信息的流向进行溯源。当系统的物理过程数据出现异常时，本文算法能够通过单向无回路的因果网络追溯异常源节点，而 BAS-TE 算法的结果中存在类似“$Stream4 \rightarrow level \rightarrow Stream10 \rightarrow Stream4$”的回路致使其溯源陷入死循环。

3.2 准确性分析

文献[20]将其实验结果与工艺流程进行了对比，而本算法则从信息熵的角度挖掘变量之间的因果控制关系，因此文本将实验结果与各器件间的实际因果逻辑进行对比，分析算法结果真实性。本文将文献[11]的实验节点 *Stream 4*、*Stream 10*、*Stream 11* 和 *level* 作为数据集 A，文献[20]的实验节点 *Stream 6*、*Stream 8*、*Stream 10*、*Stream 11*、*Valve 1* 和 *Valve 9*

作为数据集 B，其真实因果逻辑分别如图 6(a)和图 7(a)所示。各算法在两数据集合上的因果关系网络如图 6、7 所示。相对于实际因果关系，虚线为算法求得的间接因果关系。为便于分析，本文采用文献[11]中的对比方式，仅讨论实线所代表的直接因果关系。

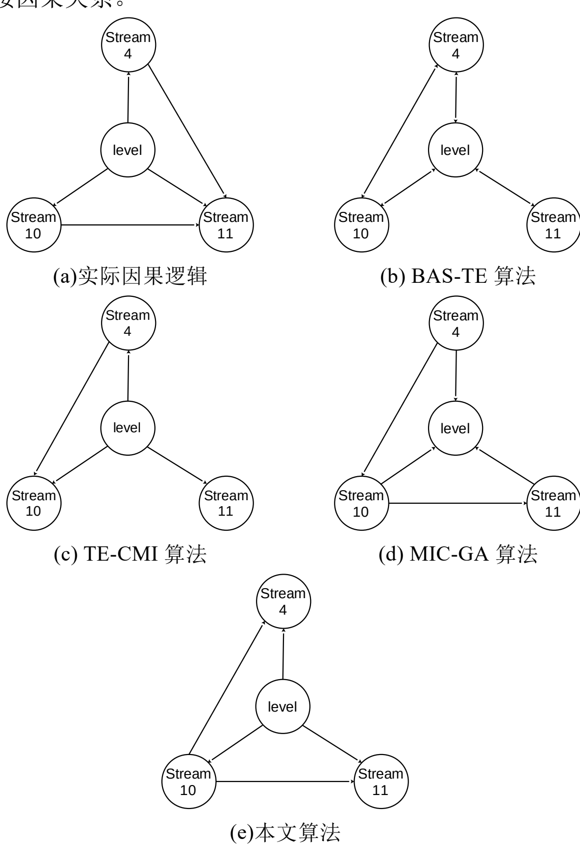


图 6 基于数据集 A 的结果对比图

Fig. 6 Results comparison diagram based on dataset A

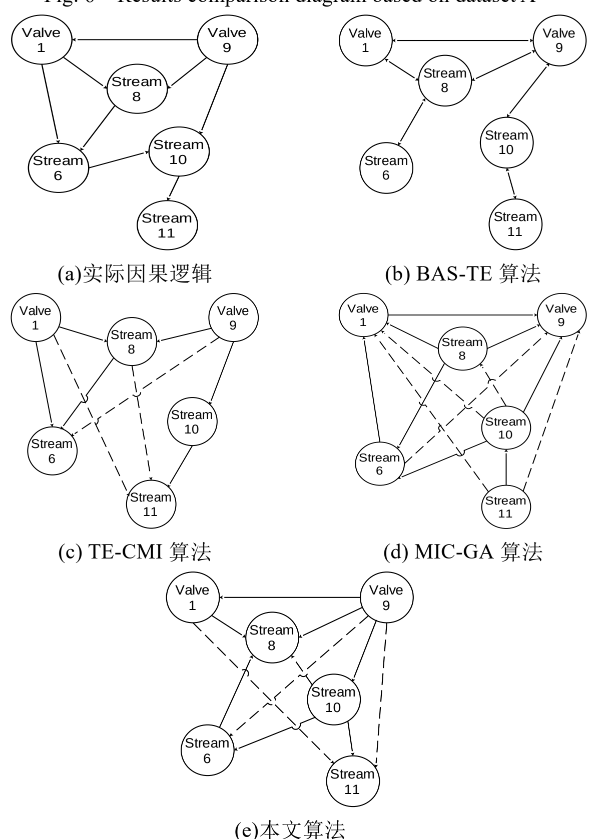


图 7 基于数据集 B 的结果对比图

Fig. 7 Results comparison diagram based on dataset B

文献[20]将算法所得结果的准确率(accuracy)视为其真实性，即

$$A=\frac{c}{s} \tag{3}$$

其中，A 表示准确率；c 表示与实际因果逻辑中相同的边数；s 表示全部可能的边数。

本文根据该真实性评判标准得到各项分类衡量指标，如表 5 所示，本文算法在准确率上明显优于 BAS-TE 算法[11]，略低于 TE-CMI 算法[20]，整体水平较好。本文算法得出的因果关系根据信息流向确定，而 MIC-GA 算法[9]则依据局部最优解确定因果方向，缺乏信息熵依据，准确率低，可靠性相对较差。

表 5 各项分类衡量指标对比

Tab. 5 Comparison of various classification measures

算法	准确率	召回率	精确率	F1	数据
BAS-TE	0.4167	0.6000	0.3750	0.4615	A
MIC-GA	0.3000	0.2000	0.2000	0.2000	A
TE-CMI	0.7500	0.6000	0.7500	0.6667	A
本算法	0.8333	0.8000	0.8000	0.8000	A
BAS-TE	0.7333	0.7500	0.5000	0.6000	B
MIC-GA	0.5333	0.1250	0.1250	0.1250	B
TE-CMI	0.9333	0.7500	1.0000	0.8571	B
本算法	0.8333	0.6250	0.7143	0.6667	B

3.3 稳定性分析

本算法通过预设预测序列长度 *len* 减少计算量，却也带来了一定程度的不稳定性。多次实验表明，预设长度 *len* 与输入序列 X、Y 的长度 L 的比值影响算法结果的稳定性。因此，本文设计了稳定性指标 S 以反映算法结果的稳定性，其计算公式为

$$S=1-\frac{\sum_{i}^{M}\frac{\Delta x}{M}}{E}\times 100\% \tag{4}$$

式(4)中，S 为稳定性比值；Δx_i 为第 i 次结果与第一次结果中的不同边数，即变化边数；M 为重复实验的次数；E 为网络中的单向有效边数，即 $\boldsymbol{mic}_{i,j}>th$ 的总数。

当本算法的 len/L 比值分别近似为 0.5、0.75、0.8、1 时，在不同维度(参与计算的节点数)的数据集下的稳定性折线图如图 8 所示，测试数据为前 3 万条数据，步长为 60(L=1: 60: 30000)。

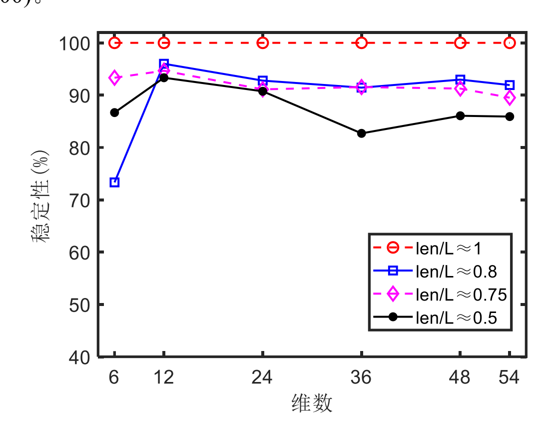


图 8 不同 len/L 取值的稳定性折线图

Fig. 8 The stability line graph of different len/L values

随着 *len/L* 的增大，算法稳定性趋于恒定。但由于 6 维数据的有效边较少，稳定性偶尔较差。由于 MIC-GA 算法[9]的同样是以 MIC 参数为基础构造因果网络，本文选取该比值的最低限度(len/L = 0.5)与 MIC-GA 算法进行稳定性对比(MIC 阈值均设置为 th = 0.1)，结果如图 9 所示。测试数据为全部 12 万条数据，步长为 60(L = 1: 60: 120000)。

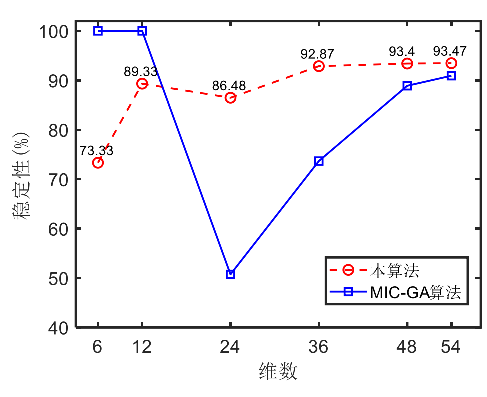


图 9　算法稳定性对比

Fig. 9　Algorithm stability comparison

MIC-GA 算法结果在维数较低时稳定性极高，但当数据维度增大后，其稳定性大幅降低，后逐渐回升。由于维数为 6 时的有效边很少(仅 3 条边)，本算法平均 0.8 的变化边数占有效边比重相对较大。但随着维数的增加，算法稳定性增强且在 24 维以后均高于对比算法，总体稳定性更好。

4　结束语

考虑到 ICS 物理过程数据的非线性以及其中包含的大量噪声数据等因素，本文提出了基于最大信息传递熵的因果关系建模算法 MITE-CM。算法利用 MIC 检测系统时序趋势之间的相关性强弱，构造接近于理论逻辑的初始网络结构，并结合传递熵 TE 判定网络间的信息流动方向形成因果关系网络。实验结果表明，本文算法可敏锐的捕捉到 ICS 物理过程数据中的异常因果关系，其结果的准确率比 BAS-TE 算法结果高出 35.74%，具有较好的真实性；在 24 维及以上的高维数据中，算法所得模型的稳定性均高于相同数据集下的 MIC-GA 算法。但本文算法整体的计算复杂度较高，后期工作将主要针对 MIC 的算法进行优化，以提高算法整体计算效率。

参考文献：

[1] Zhou Wer, Jia Yan, Peng Anni, *et al.* The effect of iot new features on security and privacy: New threats, existing solutions, and challenges yet to be solved [J]. IEEE Internet of Things Journal, 2018, 6 (2): 1606-1616.

[2] Bauer M, Cox J W, Caveness M H, *et al.* Finding the Direction of Disturbance Propagation in a Chemical Process Using Transfer Entropy [J]. IEEE Trans on Control Systems Technology, 2007, 15 (1): 12-21.

[3] Yang F, Xiao D. Progress in Root Cause and Fault Propagation Analysis of Large-Scale Industrial Processes [J]. Journal of Control Science and Engineering, 2012, 2012: 1-10.

[4] Ma Liang, Dong Jie, Peng Kaixiang. A novel key performance indicator oriented hierarchical monitoring and propagation path identification framework for complex industrial processes [J]. ISA transactions, 2019: 1275-1279.

[5] Kathari S, Tangirala A K. Efficient Reconstruction of Granger-Causal Networks in Linear Multivariable Dynamical Processes [J]. Industrial & Engineering Chemistry Research, 2019, 58 (26): 11275-11294.

[6] Granger C W J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods [J]. Econometrica, 1969, 37 (3): 424-438.

[7] Zhang Qi, Zhou Chunjie, Tian Yuchu, *et al.* A fuzzy probability bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems [J]. IEEE Trans on Industrial Informatics, 2017, 14 (6): 2497-2506.

[8] Yang Jing, Cao Jiajian. Application of continuous Bayesian network model in cross-sectional survey data. Computer Engineering and Applications, 2014, 50 (19): 192-198.

[9] 曾千千, 曾安, 潘丹, 等. 基于最大信息系数的贝叶斯网络结构学习算法 [J]. 计算机工程, 2017, 43 (8): 225-230. (Zeng Qianqian, Zeng An, Pan Dan, *et al.* Bayesian network structure learning algorithm based on maximal information coefficient [J]. Computer Engineering, 2017, 43 (8): 225-230.)

[10] Lindner B, Auret L, Bauer M, *et al.* Comparative analysis of Granger causality and transfer entropy to present a decision flow for the application of oscillation diagnosis [J]. Journal of Process Control, 2019, 79: 72-84.

[11] Shi Dawei, Guo Ziyang, Johansson K H, *et al.* Causality countermeasures for anomaly detection in cyber-physical systems [J]. IEEE Trans on Automatic Control, 2017, 63 (2): 386-401.

[12] Su Jianjun, Wang Dezheng, Zhang Yinong, *et al.* Capturing causality for fault diagnosis based on multi-valued alarm series using transfer entropy [J]. Entropy, 2017, 19 (12): 663.

[13] 麦桂珍, 彭世国, 洪英汉, 等. 混合加噪声模型与条件独立性检测的因果方向推断算法 [J]. 计算机应用研究, 2019, 36 (06): 1688-1692. (Mai Guizhen, Peng Shiguo, Hong YingHan, *et al.* Causation inference based on combining additive noise model and conditional independence [J]. Application Research of Computers, 2019, 36 (06): 1688-1692.)

[14] Downs J J, Vogel E F. A plant-wide industrial process control problem [J]. Computers & chemical engineering, 1993, 17 (3): 245-255.

[15] Capaci F, Vanhatalo E, Kulahci M, *et al.* The revised Tennessee Eastman process simulator as testbed for SPC and DoE methods [J]. Quality Engineering, 2019, 31 (2): 212-229.

[16] Liu Kangling, Fei Zhengshun, Yue Boxuan, *et al.* Adaptive sparse principal component analysis for enhanced process monitoring and fault isolation [J]. Chemometrics and Intelligent Laboratory Systems, 2015, 146: 426-436.

[17] Vanhatalo E, Kulahci M, Bergquist B. On the structure of dynamic principal component analysis used in statistical process monitoring [J]. Chemometrics and Intelligent Laboratory Systems, 2017, 167: 1-11.

[18] Kaspersky. TEP59 [EB/OL]. [2019-7-12]. https://box. kaspersky. com/d/56ad570fca/

[19] Reinartz C, Kirchhübel D, Ravn O, *et al.* Generation of Signed Directed Graphs Using Functional Models [J]. IFAC-PapersOnLine, 2019, 52 (11): 37-42.

[20] Yu Weijun, Yang Fan. Detection of Causality between Process Variables Based on Industrial Alarm Data Using Transfer Entropy [J]. Entropy, 2015, 17 (12): 5868-5887.